面向图文匹配任务的多层次图像特征融合算法 *

郝志峰 [1,2]，李俊峰 [1†]，蔡瑞初 [1]，温 雯 [1]，王丽娟 [1]，黎伊婷 [1]

(1. 广东工业大学 计算机学院，广州 510006; 2. 佛山科学技术学院 数学与大数据学院，广东 佛山 528000)

摘 要：现有主流的利用预训练卷积神经网络提取图像特征的方法存在如下问题：仅使用单层预训练特征表征图像；预训练任务与实际研究任务不一致。使得现有图文匹配方法无法充分利用图像特征，极易受到噪声特征干扰。针对上述问题，使用了预训练网络中的多层特征，并提出了多层次图像特征融合算法。在图文匹配的学习目标指导下，利用多层感知机（Multi-Layer Perceptron）有监督地融合和降维多层次的预训练图像特征，生成融合图像特征，从而充分利用预训练特征，减少噪声干扰。实验结果表明，提出的融合算法可实现对预训练的图像特征更有效的利用，相比于使用单层次特征的方法能获得更好的图文匹配效果。

关键词：图文匹配；多层次图像特征；预训练特征；融合图像特征；推荐系统

中图分类号：TP391.41　　doi: 10.19734/j.issn.1001-3695.2018.10.0780

Fusion of multi-level image features for image-text matching

Hao Zhifeng[1, 2], Li Junfeng[1†], Cai Ruichu[1], Wen Wen[1], Wang Lijuan[1], Li Yiting[1]

(1. *College of Computer*, *Guangdong University of Technology*, *Guangzhou* 510006, *China*; 2. *College of Mathematics & Big Data*, *Foshan University*, *Foshan Guangdong* 528000, *China*)

Abstract: The existing mainstream methods use the pre-trained convolutional neural networks to extract image features and usually have the following limitations: a）Only using a single layer of pre-trained features to represent image; b）Inconsistency between the pre-trained task and the actual research task. These limitations result in that the existing methods of image-text matching cannot make full use of image features and is easily influenced by the noises. To solve the above limitations, this paper used multi-layer features from a pre-trained network and proposed a fusion algorithm of multi-level image features accordingly. Under the guidance of the image-text matching objective function, the proposed algorithm fused the multi-level pre-trained image features and reduced their dimensionality using a multi-layer perceptron to generate fusion features. It is able to make full use of pre-trained features and successfully reduce the influences of noises. The experimental results show that the proposed fusion algorithm makes better use of pre-trained image features and outperforms the methods using single-level features in the image-text matching task.

Key words: image-text matching; multi-level image features; pre-trained features; Fusion features; recommendation system

0 引言

近年来，图文匹配任务在人工智能、机器学习等领域中逐渐变得热门。为了给文本选取最适合的图像，在过去通常采用人工搜索的方式，根据文本内容在海量图像中进行筛选，这会耗费人类大量的时间和精力。得益于前人所取得的成果，本文现在可以利用机器学习等技术，构建一个能根据文本内容推荐合适图像的图文匹配系统。这使得无须再进行繁琐的、重复的人工搜索，减轻工作压力。而作为一个图文匹配系统，其必须同时关注文本和图像这两个属于不同模态的研究对象，因此图文匹配实际上是属于多模态（multimodal）的任务。为了完成这个任务，一般需要解决的有三个基本问题：如何对文本进行表征；如何对图像进行表征；如何联合地分析文本和图像的特征，精准地度量两者的相似性。其中前两个表征问题尤为重要，因为它们是解决第三个问题的基础。数十年来，解决表征问题需要仔细的工程设计和相当的领域专业知识来设计一个特征提取器，将原始数据（如未处理的文本或者图像的像素值）转换成合适的内部表示或者特征向量。如此一来，建模过程会过于复杂而且往往其表征能力也不强。

基于前人在深度学习领域取得的瞩目成果，可以利用一些通用的人工神经网络去进行表征学习，例如多层感知机（multi-layer perceptron）[1]、循环神经网络（recurrent neural networks，RNN）[2]、卷积神经网络（convolutional neural networks,CNN）[3]和长短期记忆网络（long short-term memory,LSTM）[4]。这些网络是由多个简单、非线性的特征层组合而成。每个特征层都将某一级别的特征变换为更抽象、更高级的特征。有了足够的变换组合，网络也就能学习到十分复杂的功能。最关键的一点是，这些人工神经网络里的特征层并不是由人类工程师所设计的，而是在学习目标指导下从数据中学习的。因此，深度学习方法的利用简化了建模的

收稿日期：2018-10-09；**修回日期：**2018-11-18　　**基金项目：**NSFC-广东联合基金资助项目（U1501254）；国家自然科学基金资助项目（61472089）；广东省自然科学基金资助项目（2014A030306004，2014A030308008）；广东省科技计划资助项目（2015B010108006，2015B010131015）；广东省特支计划资助项目（2015TQ01X140）；广州市珠江科技新星项目（201610010101）；广州市科技计划资助项目（201604016075）

作者简介：郝志峰（1968-），男，江苏苏州人，教授，博士，主要研究方向为机器学习、人工智能等；李俊峰（1995-），男（通信作者），广东佛山人，硕士研究生，主要研究方向为机器学习、深度学习等（jefferyljf@163.com）；蔡瑞初（1983-），男，浙江温州人，教授，博士，主要研究方向为机器学习、数据挖掘等；温雯（1981-），女，江西赣州人，副教授，博士，主要研究方向为支持向量机、模式识别等；王丽娟（1978-），女，河北邢台人，副教授，博士，主要研究方向为数据挖掘、机器学习等；黎伊婷（1995-），女，广东汕尾人，硕士研究生，主要研究方向为机器学习、数据挖掘等.

过程和增强了对研究对象的表征能力。

一般地，在深度学习中为了对研究对象进行特征抽取，有两种方法：a）在研究任务的学习目标指导下有监督地训练一个人工神经网络，然后利用该网络为研究对象抽取对任务有用的特征；b）利用数据集质量较高的预训练任务训练一个人工神经网络，再用该网络中某一层特征作为研究对象的一般特征。对于一些数据集质量不够高的研究任务，为了更丰富和更有效率地对研究对象进行表征，主流的做法是采用第二种方法。例如，为了更好地抽取图像特征，可以在图像识别的任务指导下使用 ImageNet 数据集预训练一个卷积神经网络，然后使用该网络中的某一层特征层（一般是分类输出前的全连接层）的输出值作为图像特征，然后再进行进一步的研究。

人工神经网络的层级结构天然地决定了高层特征是底层特征的归纳和总结。也即网络中的不同特征层分别代表着不同层次的特征，并且随着网络层级越深，所表达的特征就越抽象和越高层次。在学习的过程中，网络必定会在任务的学习目标指导下，有监督地归纳出对任务有用的特征。然而，基于深度学习的图文匹配一般是直接使用预训练网络中的单层特征去作为图像特征，或者对该单层特征进一步进行微调（fine-tuning）。因此也就只能使用到预训练任务所归纳的某一单层次特征，或者只能从该单层次特征的基础上进一步进行归纳。遗憾的是，预训练任务和实际研究的图文匹配任务是有一定差别的（任务的不一致性）。直接使用某一单层次的预训练特征会存在图文匹配所需要的特征并没有被归纳到的情况，同时也存在大量没有作用的噪声特征；再者，对单层次的预训练特征进行微调也未能利用到其他层次的有用特征。因此，直接使用或微调预训练网络的某一单层次特征并没有充分地、合理地使用这种预训练特征，需要去抽取多层次的预训练特征，并在该多层次特征的基础上进行进一步的归纳、提炼。

特别地，针对以上问题，本文创新地使用了预训练网络中的多层特征，并提出了一种多层次图像特征融合算法（简称融合算法）。该算法通过在图文匹配任务的学习目标指导下，利用多层感知机有监督地去融合和降维多层次的预训练图像特征，最后生成出融合图像特征。其中，多层次的预训练图像特征的使用可充分地利用到更多不同层次的特征，融合和降维的过程则能归纳出对图文匹配任务有用的特征，去除无用的特征，因此也减少了噪声特征的干扰。本文之所以采用多层感知机来实现融合，是因为多层次的预训练图像特征不能简单地进行叠加融合，具有复杂的非线性关系。而多层感知机是感知机的推广，能有效地对这种非线性关系的特征进行处理，因此用多层感知机有监督地处理这种多层次的特征是一种简洁且有效的方法。

通过把融合图像特征引入到本文所实现的基于文本内容的图像推荐算法中，能够获得更好的推荐效果。最后，两个数据集上的实验结果都表明：本文所提出的方法的确能更有效地利用预训练的图像特征，生成在图文匹配任务中表达能力更强的融合图像特征。

本文的贡献主要包括以下几个方面：a）使用了由预训练卷积神经网络抽取出的多层次图像特征；b）构建了一个在图文匹配任务的学习目标指导下，对多层次的预训练图像特征进行融合和降维的多层感知机，并用其为图像生成融合图像特征；c）利用协同过滤[5]的思想构建了一个能根据文本内容进行推荐的图像推荐算法，并在该算法中使用本文所提出的融合图像特征。

1　相关工作

图文匹配在推荐系统、机器学习等领域中占据着重要的地位。Yan 等人[6]提出使用深度网络去表征图像和文本，然后利用带有深度典型关联分析的联合隐藏空间学习以解决图文匹配的问题。Ma 等人[7]在图文匹配任务中，构建了一个图像特征抽取网络，并提出使用预训练的卷积神经网络来初始化该特征抽取网络。Wang 等人[8]基于深度学习方法构建了一个图文联合隐藏空间学习的一般框架，且提出了图像和文本都存在各自的结构保持约束以及图文匹配的双向排名约束。Nam 等人[9]提出使用注意力机制去解决图文匹配以及基于视觉的问答这两种多模态任务，最后在标准的数据集中获得了十分先进的结果。

而在图像特征工程领域里，一些用于处理图像的卷积神经网络结构[10~16]变得越来越深，朝着模块化的方向发展。这些网络不断刷新着图像识别任务的成绩，现今已经能达到很高的水平，甚至已经超越了人类的识别能力。因此可以确信的是，这些优秀的网络模型有能力抽取到大量高级的图像语义特征，利用这些预训练的网络抽取图像的特征信息也是合乎常理的。

基于这些优秀的图像识别网络，Zeiler 等人[17]尝试去可视化和理解卷积神经网络，以及观察了给定的特征图是受输入图像的哪些结构所影响的。随后，Garcia-Gasulla 等人[18]尝试无监督地抽取关于抽象语义的视觉表达特征。他们使用了预训练的卷积神经网络中的卷积层特征去表征图像，然后利用与 Word-Net 距离的相关性评估了该特征空间的语义，发现该空间的向量距离是与语言语义强相关的。接着通过聚类实验，他们发现在 WordNet 中靠近的元素能被聚集在一起，“犬类”和“轮式车辆”的类别之间也存在着明显的鸿沟，而且“生物”和“非生物”这两个更高级的语义类别也能被明显地区分开来。这些证据都可以证明该表征方式的确能够成功地获取视觉的高级语义信息。Agrawal 等人[19]分析了预训练特征对于实际研究的物体识别任务是否有效。实验证明在大多数情况下，预训练特征和微调后的特征的表现都要比重新开始训练的特征要好（除了在数据集有较大补充的情况下，重新训练的特征的实验表现会比预训练特征好，但是微调特征的表现依然是最好的）。

实际上在应用研究中，大量的实验表明利用预训练的图像特征能够取得很好的效果。如 Vinyals 等人[20]使用预训练的卷积神经网络去表征图像，然后构建出一个能为图像生成描述文字的模型。Peng 等人[21]利用了微调后的预训练卷积神经网络多尺度地去获取图像的特征，并且提出了标签继承的概念。Liu 等人[22]在图像识别任务中使用了预训练卷积神经网络中的卷积层来获取图像的局部特征。由于卷积层保留了空间的信息，不再需要多次地使用网络获取图像的局部特征，因此也消除了训练图像的尺度与图像局部的尺度不一致的影响。该工作证明了只要使用得当，不仅预训练的全连接层特征有用，预训练的卷积层特征也能蕴涵十分有用的信息。Doersch 等人[23]利用了图像的自身信息，自监督地预训练网络。这个过程没有用到实际研究数据集以外的任何标签。该工作的实验结果表明，利用该方法预训练的特征也能够在计算机视觉任务中发挥作用，提升表现。以上的工作都是使用预训练的网络去表征图像的，虽然能有很好的表现，但是都仅仅使用了预训练网络中的某一单层次特征去作为图像的特征。这会导致预训练特征没有被充分地利用和易受噪声特征的干扰等问题。因此有人开始尝试去使用和融合预训练网络

中的多层次特征。如 Gatys 等人[24]在图像的风格迁移任务中使用了预训练网络中的某一单层卷积层特征来作为图像的内容特征，以及使用了由多层卷积层特征所产生的多个 Gram Matrices 共同地表示图像的风格特征。然后利用反向梯度传播修改噪声图像，最终即可产生带有指定风格和指定内容的图像。该工作通过损失函数叠加的方式对多层次的预训练特征进行利用。Liu 等人[25]利用 BoVW（bag of visual words）算法把预训练的卷积层特征转换成单词直方图以表征图像。最后通过各层直方图距离核的带权加和得到一个总的深度金字塔式匹配核，用于对 SVM（support vector machine）算法进行优化。Ronneberger 等人[26]构建了一个 U 型网络以解决生物医学图像的分割任务。该网络为了更好地进行定位，把网络的收缩路径中的高分辨率特征和上采样路径中的输出结合在一起，以致随后的连续卷积层能学习根据这些信息组合出一个更精确的输出。

2　多层次图像特征融合算法及图像推荐算法

2.1　多层次图像特征融合算法

融合算法用于对多层次的预训练图像特征进行融合和降维，是本文的核心算法，整个算法框架如图 1 所示。

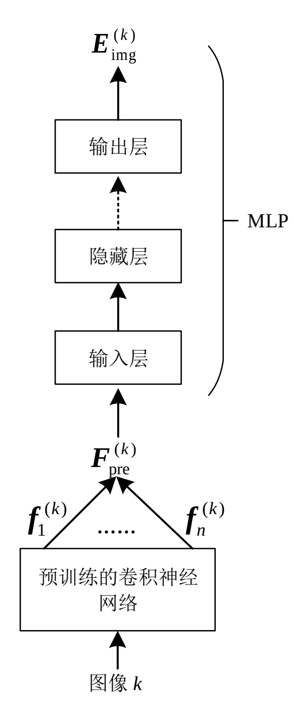


图 1　融合算法框架

Fig. 1　Framework of fusion algorithm

给定一个已经用预训练任务（如 ImageNet 图像识别任务）训练过的卷积神经网络，可以有选择地抽取使用该网络中的卷积层或全连接层的特征。假设预训练网络中共有 n 层特征层，把图像 k 输入到网络后，对各层特征进行拼接，生成一个多层次的总预训练特征 $\boldsymbol{F}_{\text{pre}}^{(k)}$：

$$\boldsymbol{F}_{\text{pre}}^{(k)}=[\boldsymbol{f}_1^{(k)},\boldsymbol{f}_2^{(k)},\ldots,\boldsymbol{f}_n^{(k)}] \tag{1}$$

其中：$\boldsymbol{f}_i^{(k)}$ 为图像 k 在预训练网络中的第 i 层特征。

为了令各层的特征能拼接在一起，当预训练网络的第 i 层特征为卷积层特征 $\boldsymbol{conv}_i^{(k)}$ 时，由于其具有空间的信息，需要对该层特征进行池化（Pooling）操作以消除空间信息：

$$\boldsymbol{f}_i^{(k)}=\text{pool}(\boldsymbol{conv}_i^{(k)}) \tag{2}$$

而当第 i 层特征为全连接层特征 $\boldsymbol{fc}_i^{(k)}$ 时，由于其不具有空间信息，则不需要进行池化操作：

$$\boldsymbol{f}_i^{(k)}=\boldsymbol{fc}_i^{(k)} \tag{3}$$

需要注意的是，实际上不总是需要使用预训练网络中所有的特征层，而是可以根据具体情况有选择性地使用，因此式（1）中的 $\boldsymbol{F}_{\text{pre}}^{(k)}$ 可以只包含部分特征层的特征。

为了从多层次的预训练特征中归纳出对图文匹配任务有用的特征和舍弃无用的噪声特征，本文构建了一个多层感知机 MLP 在图文匹配任务的学习目标指导下去融合和降维图像 k 的 $\boldsymbol{F}_{\text{pre}}^{(k)}$ 特征，最终输出融合图像特征 $\boldsymbol{E}_{\text{img}}^{(k)}$：

$$\boldsymbol{E}_{\text{img}}^{(k)}=\text{MLP}(\boldsymbol{F}_{\text{pre}}^{(k)}) \tag{4}$$

该 MLP 为标准的全连接人工神经网络，其设计有以下几个特点：a）隐藏层和输出层都设置了非线性激活函数以增强网络的表达能力；b）网络的各层维度随着深度越深变得越低，用于对高维度且包含大量噪声特征的多层次预训练特征进行融合和降维；c）网络输出的融合图像特征的维度要与文本特征一致，以便进行相似度测量。由于 MLP 中所采用的激活函数、各层的维度以及层数是与实际的研究对象和研究数据集有关的，本节并没有更详细地定义其网络的细节结构，只是给出一个最一般的定义，实际所用 MLP 的更多细节将在第 3 章中呈现。为了训练 MLP 的网络参数，定义一个约束：

$$s(\boldsymbol{E}_{\text{text}}^{(x_i)},\boldsymbol{E}_{\text{img}}^{(y_j)})>m+s(\boldsymbol{E}_{\text{text}}^{(x_i)},\boldsymbol{E}_{\text{img}}^{(y_p)}) \\ \forall y_j\in Y_i^+,\forall y_p\in Y_i^- \tag{5}$$

其中：Y_i^+ 和 Y_i^- 分别代表训练文本 x_i 所对应的正类（匹配）和负类（不匹配）的图像集合；$\boldsymbol{E}_{\text{text}}^{(x_i)}$ 为文本 x_i 对应的特征向量（通过一些无监督方法，如潜语义分析（latent semantic analysis，LSA）主题模型[27]和 doc2vec[28]等，去抽取出文本特征向量）；$\boldsymbol{E}_{\text{img}}^{(y_j)}$ 和 $\boldsymbol{E}_{\text{img}}^{(y_p)}$ 分别代表当图像 y_j 和 y_p 作为融合算法的输入时，所输出的融合图像特征；$s(\boldsymbol{v}_1,\boldsymbol{v}_2)$ 代表着 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 的余弦相似度；$\boldsymbol{m}$ 为强制间隔大小（Enforced Margin）。

式（5）的约束表示：给定训练文本 x_i，令其与对应的每个正类图像 y_j 的特征相似度，都要大于间隔大小 $\boldsymbol{m}$ 加上其与每个负类图像 y_p 的特征相似度。

通过使用 Hinge Loss 的标准形式，把（5）的约束转换为 MLP 的训练损失函数：

$$\text{loss}=\sum_{i,j,p}\max[0,m+s(\boldsymbol{E}_{\text{text}}^{(x_i)},\boldsymbol{E}_{\text{img}}^{(y_p)})-s(\boldsymbol{E}_{\text{text}}^{(x_i)},\boldsymbol{E}_{\text{img}}^{(y_j)})] \tag{6}$$

式（6）的损失函数包含了训练集中所有由训练文本，对应的正类图像，以及对应的负类图像所组成的三元组。而由于三元组的组合数量太多，使用所有的三元组来训练 MLP 是不切实际的。所以在 MLP 的每一次迭代训练中，对于每个训练文本，仅随机选取一个负类图像，和对应的正类图像共同构建出三元组以进行匹配训练。

实际上，对于不同的训练样例，式（6）中的间隔大小 $\boldsymbol{m}$ 是可以不同的。但是为了更易于进行优化，为数据集中的所有训练样例设置一个固定的间隔大小 $\boldsymbol{m}$，其具体数值将在章节 3 中给出。

2.2　图像推荐算法

在本节实现了一个基于协同过滤思想的、能根据文本内容进行推荐的图像推荐算法，并且在搜狐 2017 图文匹配大赛（https://www.biendata.com/competition/luckydata/）中使用了该算法，获得了第三名。设 X_{train} 代表训练集中的文本集合，Y_{test} 代表测试集中的图像集合以及 result 代表推荐结果的图像集

合。算法的具体步骤为:

a）给定一个在测试集中的文本 x_{test}，通过文本主题模型为其在文本集合 X_{train} 中找寻最相似文本内容的文本 x^*_{train}（式（7）），相应地，即可获得文本 x^*_{train} 在训练集中所对应的匹配图像 y^*_{train}；

b）在图像集合 $(Y_{\text{test}}-result)$ 中利用图像的特征信息找寻出与图像 y^*_{train} 最相似的图像 y^*_{test} 作为推荐候选图像（式（8）），并把该 y^*_{test} 放进图像集合 *result* 中。

$$x^*_{\text{train}}=\arg\max_{x^i_{\text{train}}\in X_{\text{train}}} s(\boldsymbol{E}^{(x_{\text{test}})}_{\text{text}},\boldsymbol{E}^{(x^i_{\text{train}})}_{\text{text}}) \tag{7}$$

$$y^*_{\text{test}}=\arg\max_{y^j_{\text{test}}\in(Y_{\text{test}}-result)} s(\boldsymbol{E}^{(y^*_{\text{train}})}_{\text{img}},\boldsymbol{E}^{(y^j_{\text{test}})}_{\text{img}}) \tag{8}$$

该推荐流程如图 2 所示。重复步骤 b）直至 *result* 中已包含 *K* 个推荐候选图像。

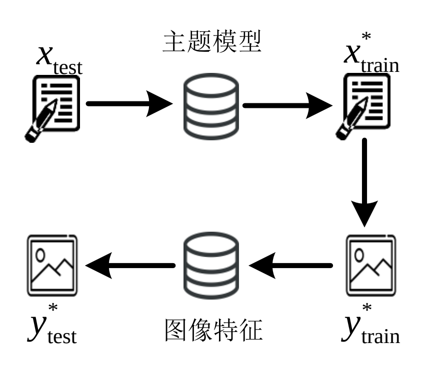


图 2　推荐流程图

Fig. 2　Recommendation flowchart

显然，在本推荐算法中使用不同表征能力的图像特征会产生不同的推荐表现。本文尝试把多种图像特征（包括由本文 2.1 节的融合算法所生成的融合图像特征）分别作为推荐算法中的图像特征信息，然后根据推荐表现直接地评估各种图像特征的表征能力。

3　实验及结果分析

本章在搜狐图文匹配比赛数据集和 Flickr30K 数据集[29]上进行了对比实验，评估了各种图像特征的表现，并结合实验结果分析了本文所提出的融合算法的优势。本文使用图文匹配任务中常用的评测标准 recall@*K*/%（*K*=1, 5, 10）报告了实验的推荐表现，其为匹配的图像被检索在推荐结果前 *K* 的新闻文本占得的比例。

3.1　搜狐图文匹配比赛数据集

本数据集来源于 2017 年由搜狐公司举办的图文匹配大赛。本实验所使用的数据包含了初赛十万级别的训练集，复赛百万级别的训练集，决赛 400 小型测试集（决赛 400 小型测试集是在决赛 20000 完整测试集中分出的子集，包含了完整测试集中的 400 篇新闻和对应 400 幅配图）以及决赛 20000 完整测试集。该数据集里的每一篇新闻文本都有其相应的一幅配图。

在本实验中，为了更好地表达文本，利用百万组数据级别的复赛训练集中的所有新闻文本，训练了一个 500 主题的潜语义分析主题模型[27]，并通过该模型为所有的新闻文本生成特征向量。融合算法里的预训练网络是经过 ImageNet 图像识别任务预训练完成的 Inception v3 网络[14]，该网络的结构轮廓在表 1 给出，更详细的结构可参见文献[14]。本实验使用了预训练网络中两个特征层来作为多层次预训练图像特征，分别是：最后的 Pool 层特征（简称 fc 特征）；首个 Inception 模块 3（图 6）的经过最大值池化处理的卷积层特征输出（简称 mixed9 特征）。

表 1　Inception v3 网络结构

Table 1　Network structure of Inception v3

类型	窗口大小/步长或备注	输入大小
Conv	3 × 3/2	299 × 299 × 3
Conv	3 × 3/1	149 × 149 × 32
Conv Padded	3 × 3/1	147 × 147 × 32
Pool	3 × 3/2	147 × 147 × 64
Conv	3 × 3/1	73 × 73 × 64
Conv	3 × 3/2	71 × 71 × 80
Conv	3 × 3/1	35 × 35 × 192
3 × Inception	图 4	35 × 35 × 288
5 × Inception	图 5	17 × 17 × 768
2 × Inception	图 6	8 × 8 × 1280
Pool	8 × 8	8 × 8 × 2048
Linear	Logits	1 × 1 × 2048
Softmax	Classifier	1 × 1 × 1000

为了设定融合算法中 MLP 的结构参数，进行了以下探讨。Cybenko[30]已经证明了，MLP 最多只需要一层隐藏层就能够达到近似函数的目的。基于该结论，本文的 MLP 只设置一层隐藏层。此外还进一步探究了非线性激活函数对融合效果的影响，在图 3 呈现（利用推荐效果来间接体现融合效果）。可看到为隐藏层加入了非线性激活函数后，网络的融合效果下降了，但是通过加入对网络参数的 L2 正则化约束后，能使整个网络的拟合能力和泛化能力提升，融合效果为最优。所以本文最后采用了为隐藏层加入非线性激活函数的网络结构，并且利用 L2 正则化优化网络参数的训练。具体地，融合算法中使用了 fc 特征和 mixed9 特征作为输入的 MLP（简称 $\text{MLP}_{\text{fc,mixed9}}$）的结构为：维度为 4096 维的输入层；维度为 2048 维，带有 sigmoid 激活函数的隐藏层；维度为 500 维，带有 tanh 激活函数的输出层。本文使用了十万组数据级别的初赛训练集在最小化式（6）的指导下有监督地训练 $\text{MLP}_{\text{fc,mixed9}}$ 的网络参数，并且在训练的过程中加入了对参数的 L2 正则化约束（权重为 0.0005）以优化训练效果。其中，式（6）损失函数中的 *m* 设置为 0.5，参数采用 Adam 算法（learning_rate=0.001，beta1=0.9，beta2=0.999，e-psilon=1e-08）进行更新。

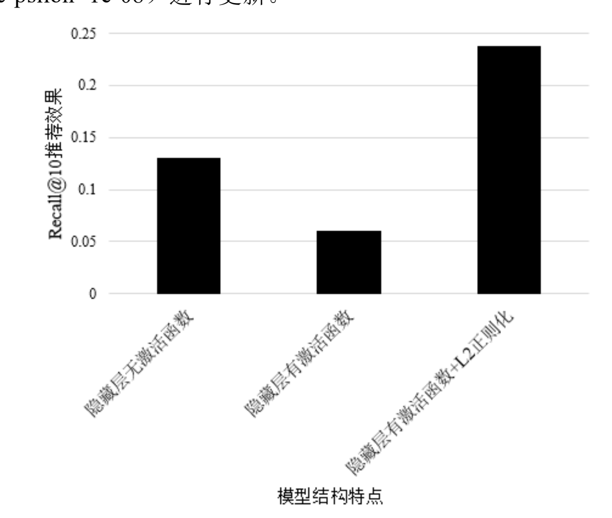


图 3　MLP 网络结构对融合效果的影响

Fig. 3　Influence of MLP network structure on fusion performance

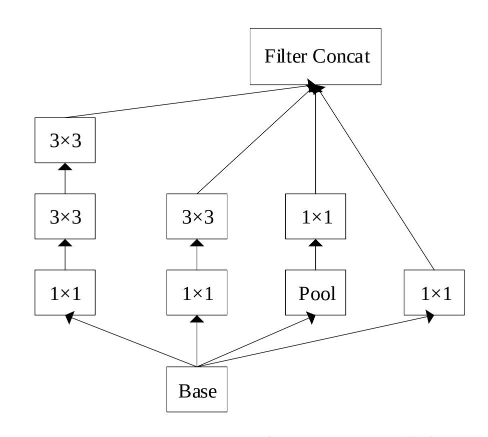


图 4　Inception v3 网络中的 Inception 模块 1

Fig. 4　Module 1 in Inception v3 network

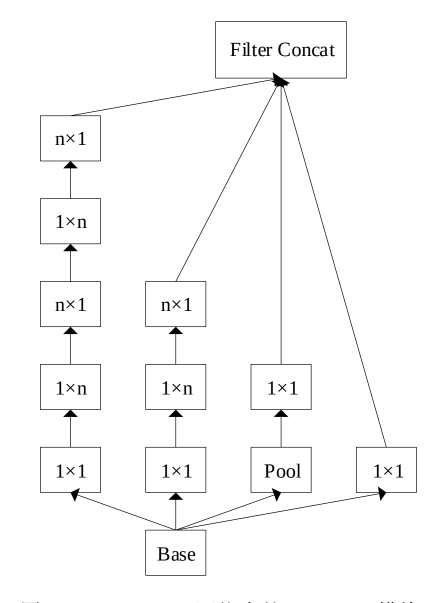


图 5　Inception v3 网络中的 Inception 模块 2

Fig. 5　Module 2 in Inception v3 network

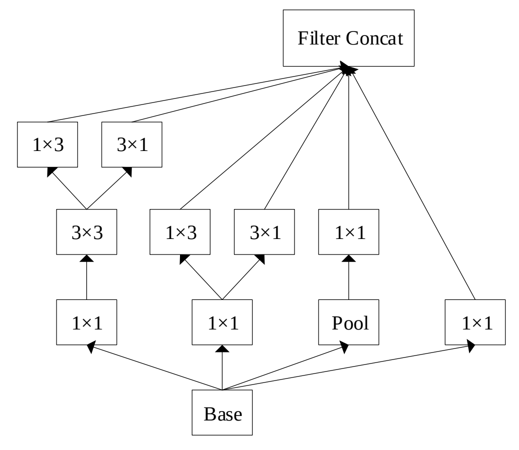


图 6　Inception v3 网络中的 Inception 模块 3

Fig. 6　Module 3 in Inception v3 network

为了直接对比由 $\mathrm{MLP_{fc,mixed9}}$ 生成的融合图像特征（简称 $\mathrm{MLP_{fc,mixed9}}$ 融合图像特征）、fc 单层次特征和 fc+mixed9 多层次特征的推荐表现，本文在 2.2 节的图像推荐算法中分别使用了这三种图像特征来进行对比实验，实验结果在表 2 给出（其中，在图像推荐算法中采用 fc 单层次特征是本团队在 2017 搜狐图文匹配大赛中的做法，前两名团队的图像表征方法也是类似的）。根据在决赛 400 小型测试集的实验结果得出，fc+mixed9 特征的 Recall@K 表现相比于 fc 特征都有稍微的提高，其中 recall@10 表现有 0.5 的提升。证明多层次特征比单层次特征会有更好的推荐表现，但是由于其特征维度太高，含有大量对图文匹配任务而言是噪声的特征，所以表现提升不算太大。然而本文的方法能够在图文匹配任务的学习目标指导下，有监督地对多层次图像特征进行融合和降维。因此，$\mathrm{MLP_{fc,mixed9}}$ 融合图像特征的 recall@K 表现相比于 fc 特征都有大量的提高，其中 Recall@10 表现甚至有 9.5 的提升。显然，本文提出的方法是有效的。为了进一步检验本方法在大型测试集中是否有效，本文在决赛 20000 完整测试集中也进行了该实验，实验结果仍然在表 2 给出。从表 2 可以看出，因为推荐的搜索空间扩大了，所以完整测试集中的 recall@K 表现明显比小型测试集的要差。尽管如此，fc+mixed9 特征的大部分 Recall@K 表现还是相比于 fc 特征有稍微的提高。而 $\mathrm{MLP_{fc,mixed9}}$ 融合图像特征的 recall@K 表现相比于 fc 特征则有更大的提升，其中 recall@10 表现有 0.6 的提升。因而，本方法在大型测试集中也是有效的。

表 2　搜狐数据集中各种图像特征在图像推荐算法的对比实验

Table 2　Comparison of image recommendation algorithms using different image features on SOHU dataset

测试集	特征	R@1	R@5	R@10
决赛 400 小型测试集	fc 特征	4.0	7.0	10.2
	fc+mixed9 特征	4.7	8.0	10.7
	$\mathrm{MLP_{fc,mixed9}}$ 融合图像特征（this paper）	**6.2**	**13.0**	**19.7**
决赛 20000 完整测试集	fc 特征	0.6	1.2	1.7
	fc+mixed9 特征	**0.7**	1.3	1.7
	$\mathrm{MLP_{fc,mixed9}}$ 融合图像特征（this paper）	**0.7**	**1.6**	**2.3**

本实验还设计了一个只使用 fc 特征作为输入的 MLP（简称 $\mathrm{MLP_{fc}}$，其产生的图像特征也简称为 $\mathrm{MLP_{fc}}$ 特征），其具体结构和训练细节与 $\mathrm{MLP_{fc,mixed9}}$ 相比，只是结构上的输入层维度变为 2048，其他保持一致。由于 MLP 的训练约束能让其输出的图像特征向量与文本特征向量直接在余弦相似度上进行匹配，所以实验在表 3 分别给出了利用不同 MLP 所输出的图像特征与文本特征直接进行相似度匹配的推荐表现，以检验在融合和降维的过程中使用多层次的特征是否比使用单层次的特征更有优势。

表 3　在搜狐数据集中直接对图像特征和文本特征进行相似度匹配的对比实验

Table 3　Comparison of similarity matching between image features and text features on SOHU dataset

测试集	特征	R@1	R@5	R@10
决赛 400 小型测试集	$\mathrm{MLP_{fc}}$ 特征	**5.5**	13.7	20.7
	$\mathrm{MLP_{fc,mixed9}}$ 融合图像特征	4.2	**16.7**	**24.2**
决赛 20000 完整测试集	$\mathrm{MLP_{fc}}$ 特征	0.3	1.5	2.4
	$\mathrm{MLP_{fc,mixed9}}$ 融合图像特征	**0.5**	**1.7**	**3.0**

根据表 3 可以看到，$\mathrm{MLP_{fc,mixed9}}$ 融合图像特征的大部分 recall@K 表现要优于 $\mathrm{MLP_{fc}}$ 特征（只在小型测试集中的 recall@1 出现了 $\mathrm{MLP_{fc,mixed9}}$ 融合图像特征表现较差的情况）。该结果表明，在图文匹配任务的学习目标指导下，$\mathrm{MLP_{fc,mixed9}}$ 所产生的融合图像特征要比 $\mathrm{MLP_{fc}}$ 所产生的特征要有更强的表达能力。换言之，在融合和降维的过程中使用多层次的特征的确比使用单层次的特征更有优势。同时也能看到，与图像推荐算法相比，利用 $\mathrm{MLP_{fc,mixed9}}$ 融合图像特征直接进行相似度匹配的方法在 recall@5,10 上具有更优秀的表现，尤其是在

recall@10 表现上，其相对于只采用 fc 特征的图像推荐算法在小型测试集里有 14.0 的提升，在完整测试集里有 1.3 的提升。

3.2　Flickr30K 数据集

为了验证本文提出的融合算法是否具有普遍性，本文在 Flickr30K 数据集中也进行了对比实验。该数据集总共包含了 31783 幅图像，每幅图像都有其对应的 5 个描述短句。本实验遵循了公开的数据集划分方案[31]，把该数据集分成了 29 783 幅训练图像、1 000 幅验证图像以及 1 000 幅测试图像。

在本实验中，利用所有的描述短句训练出 500 主题的潜语义分析主题模型，用于产生文本的特征向量。与 3.1 节的实验一样，本节实验也使用了预训练 Inception v3 网络中的 fc 特征以及 mixed9 特征来作为多层次预训练图像特征。融合算法中所使用的 $MLP_{fc,mixed9}$ 和用于对比实验的 MLP_{fc} 的具体结构和训练细节也与搜狐数据集实验保持一致，已在 3.1 节给出。最终则能利用 29783 幅训练图像和对应的描述短句，有监督地对 $MLP_{fc,mixed9}$ 和 MLP_{fc} 的网络参数进行训练。

在本数据集中，各种图像特征在图像推荐算法中的对比实验结果在表 4 给出。从表 4 可以看到，与 3.1 节的实验结果不一样，fc+mixed9 特征的大部分 recall@K 表现相比于 fc 特征要差，证明噪声特征在该数据集中有很严重的不利影响。而 $MLP_{fc,mixed9}$ 融合图像特征相比于 fc 特征，尽管在 Recall@1 的表现有下降，但是其在 recall@5,10 的表现更优，其中 recall@10 表现有 1.0 的提升。因此，本文方法在 Flickr30K 数据集中也是有效的。

表 4　Flickr30K 数据集中各种图像特征在图像推荐算法的对比实验

Table 4　Comparison of image recommendation algorithms using different image features on Flickr30K dataset

测试集	特征	R@1	R@5	R@10
1000 测试集	fc 特征	4.3	13.1	19.1
	fc+mixed9 特征	**4.5**	12.6	18.5
	$MLP_{fc,mixed9}$ 融合图像特征（this paper）	3.6	**13.3**	**20.1**

同样地，本文也在 Flickr30K 数据集中对使用单层次特征和使用多层次特征的 MLP 进行对比实验，实验结果在表 5 给出。可以看出，$MLP_{fc,mixed9}$ 融合图像特征在 Recall@K 表现上全面优于 MLP_{fc} 特征。所以在 Flickr30K 数据集中，同样能得出在融合和降维的过程中使用多层次的特征比使用单层次的特征更有优势的结论。并且，相比于图像推荐算法，利用 $MLP_{fc,mixed9}$ 融合图像特征直接进行相似度匹配的方法在 recall@K 上的表现也要全面占优，尤其在 recall@10 表现上，其相对于只采用 fc 特征的图像推荐算法甚至有高达 20.6 的提升。

表 5　在 Flickr30K 数据集中直接对图像特征和文本特征进行相似度匹配的对比实验

Table 5　Comparison of similarity matching between image features and text features on Flickr30K dataset

测试集	特征	R@1	R@5	R@10
1000 测试集	MLP_{fc} 特征	8.5	25.0	36.9
	$MLP_{fc,mixed9}$ 融合图像特征	**9.5**	**27.9**	**39.7**

3.3　小结

综合两个数据集的实验结果，本文提出的融合算法的确能有效地对多层次的预训练图像特征进行融合和降维，充分地利用预训练的特征，最后生成在图文匹配任务中表达能力更强的融合图像特征。特别是由于搜狐比赛数据集是由现实世界真实存在的新闻及其配图所组成，因此在该数据集上解决图文匹配任务的难度会更大。但是本文的方法在该数据集上依然是有效的，能获得更好的推荐表现。

4　结束语

图文匹配一直是一个极具挑战性的任务，需要精准地抽取文本和图像的特征。特别是对于图像来说，由于其表达同样事物的表现更为丰富，因而获取图像特征尤为困难。大量先前的研究工作提出了多种获取图像特征的方法，而现今主流的做法是使用预训练深度学习网络去抽取图像特征。然而，该主流做法未能充分利用有用的特征且易受噪声特征的干扰。

针对以上存在的问题，本文提出了一种对多层次深度表达的预训练图像特征进行利用的融合算法：通过利用图文匹配任务的学习目标，有监督地融合和降维多层次的预训练图像特征，最终生成出融合图像特征，充分地利用了更多的有用特征和减少了噪声特征的干扰。在实验部分，通过把融合图像特征引入到本文实现的图像推荐算法中进行对比实验，证明了融合图像特征确实是拥有更强大的表征能力，能获得更好的推荐表现。而且，本文也进一步地设计了一个实验，证明了在融合和降维的过程中使用多层次的特征是比使用单层次的特征更有优势，获得的效果更好。最后综合所有的实验结果，得出本文所提出的方法是有效的结论。值得注意的是，本文虽然针对的是图文匹配任务，但实际上通过更改用于指导融合和降维的学习目标，可以把本方法延伸到不同任务中。

目前在短句文本上的图文匹配已经获得了很好的成果。但是对于长句文本来说，由于其内容十分复杂，难于抽取关键特征，在短句中适用的方法并不适用于长句。因此在长句文本上的图文匹配将是一个需克服的挑战。

参考文献：

[1] 宋宜斌. 多层感知器的一种快速网络训练法及其应用 [J]. 控制与决策, 2000, 15 (1): 125-127. (Song Yibin. Quick training method for multi-layer perception and its application [J]. Control and Decision, 2000, 15 (1): 125-127.)

[2] Lipton Z C, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning [EB/OL]. (2015-10-17) . https://arxiv. org/abs/1506. 00019.

[3] LeCun Y, Bottou L, Bengio Y, *et al.* Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86 (11): 2278-2324.

[4] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9 (8): 1735-1780.

[5] 冷亚军, 陆青, 梁昌勇. 协同过滤推荐技术综述 [J]. 模式识别与人工智能, 2014, 27 (8): 720-734. (Leng Yajun, Lu Qing, Liang Changyong. Survey of recommendation based on collaborative filtering [J]. Pattern Recognition and Artificial Intelligence, 2014, 27 (8): 720-734.)

[6] Yan Fei, Mikolajczyk K. Deep correlation for matching images and text [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2015: 3441-3450.

[7] Ma Lin, Lu Zhengdong, Shang Lifeng, *et al.* Multimodal convolutional neural networks for matching image and sentence [C]// Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society,2015: 2623-2631.

[8] Wang Liwei, Li Yin, Lazebnik S. Learning deep structure-preserving image-text embeddings [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society,2016: 5005-5013.

[9] Nam H, Ha J W, Kim J. Dual attention networks for multimodal reasoning and matching [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society,2017: 2156-2164.

[10] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012: 1097-1105.

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10) . https://arxiv. org/abs/1409. 1556.

[12] Szegedy C, Liu Wei, Jia Yangqing, *et al.* Going deeper with convolutions [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2015: 1-9.

[13] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C]// Proc of ACM International Conference on Machine Learning. New York: ACM Press, 2015: 448-456.

[14] Szegedy C, Vanhoucke V, Ioffe S, *et al.* Rethinking the inception architecture for computer vision [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society,, 2016: 2818-2826.

[15] Szegedy C, Ioffe S, Vanhoucke V, *et al.* Inception-v4, inception-resnet and the impact of residual connections on learning [C]// Proc of AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2017: 4278-4284.

[16] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Deep residual learning for image recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016: 770-778.

[17] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2014: 818-833.

[18] Garcia-Gasulla D, Ayguadé E, Labarta J, *et al.* A visual embedding for the unsupervised extraction of abstract semantics [EB/OL]. (2016-12-16) . https://arxiv. org/abs/1507. 08818.

[19] Agrawal P, Girshick R, Malik J. Analyzing the performance of multilayer neural networks for object recognition [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2014: 329-344.

[20] Vinyals O, Toshev A, Bengio S, *et al.* Show and tell: a neural image caption generator [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2015: 3156-3164.

[21] Peng Kuanchuan, Chen T. A framework of extracting multi-scale features using multiple convolutional neural networks [C]// Proc of IEEE International Conference on Multimedia and Expo. Piscataway, NJ: IEEE Press, 2015: 1-6.

[22] Liu Lingqiao, Shen Chunhua, van den Hengel A. The treasure beneath convolutional layers: cross-convolutional-layer pooling for image classification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2015: 4749-4757.

[23] Doersch C, Gupta A, Efros A A. Unsupervised visual representation learning by context prediction [C]// Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2015: 1422-1430.

[24] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016: 2414-2423.

[25] Liu Qingshan, Hang Renlong, Song Huihui, *et al.* Adaptive deep pyramid matching for remote sensing scene classification [EB/OL]. (2016-11-11) . https://arxiv. org/abs/1611. 03589.

[26] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation [C]// Proc of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234-241.

[27] Evangelopoulos N E. Latent semantic analysis [J]. Wiley Interdisciplinary Reviews: Cognitive Science, 2013, 4 (6): 683-692.

[28] Le Q, Mikolov T. Distributed representations of sentences and documents [C]// Proc of ACM International Conference on Machine Learning. New York: ACM Press, 2014: 1188-1196.

[29] Young P, Lai A, Hodosh M, *et al.* From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions [J]. Transactions of the Association for Computational Linguistics, 2014 ,2: 67-78.

[30] Cybenko G. Approximation by superpositions of a sigmoidal function [J]. Mathematics of Control, Signals and Systems, 1989, 2 (4): 303-314.

[31] Mao Junhua, Xu Wei, Yang Yi, *et al.* Deep captioning with multimodal recurrent neural networks (M-RNN) [EB/OL]. (2015-06-11) . https://arxiv. org/abs/1412. 6632.